DOCUMENT RESUME

ED 053 210 TM 000 743

AUTHOR Campbell, Joel T.; And Others
TITLE Prediction of Job Performance for Negro and White Medical Technicians: The Prediction of Supervisors' Ratings from Aptitude Tests, Using a Cross-Ethnic Cross-Validation Procedure.
INSTITUTION Educational Testing Service, Princeton, N.J.
REPORT NO PR-70-18
PUB DATE Oct 70
NOTE 25p.

EDRS PRICE EDRS Price MF-$0.65 HC-$3.29
DESCRIPTORS Aptitude Tests, Caucasians, Cross Cultural Studies, Culture Free Tests, Job Skills, *Medical Laboratory Assistants, Multiple Regression Analysis, *Negroes, Performance Criteria, Personnel Selection, *Prediction, Predictive Ability (Testing), Racial Discrimination, Rating Scales, *Supervisors, *Task Performance, Test Bias, Tests, Test Validity, Weighted Scores

ABSTRACT

This paper, a continuation of the analysis of data collected on 455 Negro and white medical technicians (see ED 035 018) as part of a study on fairness in selection testing, investigates the prediction of supervisors' ratings from aptitude tests. (AG)

ED053210

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
OFFICE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY

PR-70-18

PREDICTION OF JOB PERFORMANCE FOR NEGRO AND WHITE MEDICAL TECHNICIANS

The Prediction of Supervisors' Ratings from Aptitude Tests, Using a Cross-Ethnic Cross-Validation Procedure

Joel T. Campbell
Lewis W. Pike
Ronald L. Flaugher
and
Margaret H. Mahoney

October 1970

EDUCATIONAL TESTING SERVICE
PRINCETON, NEW JERSEY

1

ED053210

Personnel

Management Committee

Albert P. Maslow, Chief, Personnel Measurement Research and Development Center, U. S. Civil Service Commission

Samuel J. Messick, Vice-President, Educational Testing Service

Advisory Committee

John K. Hemphill, Far West Laboratory for Educational Research and Development, Chairman

Marvin D. Dunnette, University of Minnesota

Robert M. Guion, Bowling Green State University

S. O. Roberts, Fisk University

Project Staff

From U. S. Civil Service Commission

William A. Gorham, Program Director, Personnel Measurement Research and Development Center

From Educational Testing Service

Joel T. Campbell, Senior Research Psychologist, Principal Investigator

Ronald L. Flaugher, Research Psychologist

Lewis W. Pike, Research Psychologist

Donald A. Rock, Research Psychologist

Franklin R. Evans, Associate Research Psychologist

Mrs. Lila Norris, Associate Research Psychologist

Mrs. Margaret Mahoney, Senior Research Assistant

Mrs. Mary Ellen Parry, Senior Research Assistant

Mrs. Harriet Blizzard, Research Assistant

Mrs. Virginia Rau, Administrative Assistant

TABLE OF CONTENTS

	PAGE
Introduction	1
Part I - Single Tests	2
Procedure and Results	2
Discussion	3
Part II - Weighted Test Batteries	6
Procedure	6
Correlations	7
Multiple Correlation Coefficients	8
Cross-Validation Coefficients	9
Predictions Resulting from Multiple Regression Equation	10
Summary and Conclusions	11
Tables 1 through 6	13 - 18
Figures 1 through 3	19 - 21
Reference	22

Prediction of Supervisors' Ratings From Aptitude Tests, Using a Cross-Ethnic, Cross-Validation Procedure

Joel T. Campbell, Lewis, W. Pike, Ronald L. Flaugher, and Margaret H. Mahoney

INTRODUCTION

This paper continues the analysis of the data collected on 455 Medical Technicians in U.S. Veteran's Administration Hospitals across the country, as part of a study on fairness in selection testing.

A previous paper (Flaugher, Campbell, and Pike, 1969) reported that the race of both the person being rated and the person doing the rating has a noteworthy influence on the evaluation received. In particular, those factors measured by a Job Knowledge Test appeared to have a sizable influence on white supervisors' ratings of the job performance of both Negro and white technicians, and on Negro supervisors' ratings of Negroes; while Negro supervisors' ratings of white technicians appeared to be essentially unrelated to factors measured by the Job Knowledge Test. The present paper carried the implications of this finding a step further by asking the question: "Given the existence of an interaction between race of rater and race of ratee, what would be the consequences, first, if an aptitude test which is valid for one rater-ratee combination were used to select individuals in the other three combinations; and second, if a best weighted battery for one ethnic group were used to select individuals in the other three combinations?"

This type of investigation has particular importance for the study of test bias in general. In any validation attempt the ideal situation, from

---

[1]This study was funded by the Ford Foundation and conducted jointly by the Educational Testing Service and the United States Civil Service Commission.

a statistical standpoint, is that of an infinite population of applicants from which a random sample can be selected, tested, placed on the job, and evaluated. In practice, of course, this condition is seldom even roughly approximated.

Most validation studies of necessity use relatively small samples of employees who are already on the job. The present study is similarly removed from the theoretical ideal, in that it is confined to persons already employed. In this study, however, the samples of both Negro and white technicians are reasonably large, and there are both Negro and white supervisors. These facts make it possible to consider what happens when a test is validated against a criterion for one rater-ratee ethnic combination, and is then used for selection of both majority and minority group employees who will be working for supervisors who may belong either to a majority or a minority ethnic group.

This is, in fact, what is most likely to happen in the typical validity study. A selection measure is validated on present employees, in many cases predominantly white, against a criterion of job ratings by supervisors who are also predominantly white, and the selection measure validated in this manner is then used to screen applicants, including Negroes and others in minority groups.

PART I - SINGLE TESTS

Procedure and Results

Correlation coefficients were computed separately for each rater-ratee

ethnic combination between the overall supervisory ratings of technicians' job performance and each of nine aptitude tests. Results are shown in Table 1.

To answer the question of what would happen if a test which is valid for one rater-ratee combination were then applied to the other three, the test which was most valid for each combination was identified. The most valid test for whites rated by whites was the short-term memory (Picture-Number) test; that for whites rated by Negroes was the spatial visualization (Paper Folding) test; and that for both of the Negro ratee groups was the number facility (Subtraction and Multiplication) test. For each of these tests regression lines were then computed for all four rater-ratee ethnic combinations, and the slopes and positions of the lines were compared. The regression lines for each of these three tests, predicting overall ratings for each rater-ratee group, are shown in Figures 1, 2, and 3, respectively. The mean score and standard deviation intervals have been indicated on each regression line.

Discussion

In Figure 1, it can be seen that virtually any cutting score on the Picture-Number test, say 20, would effectively discriminate between white technicians rated high by white supervisors and those rated low by white supervisors. However, the cutting score would produce very little or no discrimination between high- and low-rated technicians in each of the other rater-ratee combinations.

In Figure 2, a similar observation can be made for the Spatial Visualization test which showed the highest validity for the white technicians rated by Negro supervisors. A cutting score set at 9, for example, would effectively differentiate between technicians in this group who were rated high and those rated low. For the two groups rated by white supervisors, however, such a cutting score would produce a selected group which was only slightly better on the criterion than was the rejected group; while for the group of Negro technicians rated by Negro supervisors, the technicians selected would actually be slightly lower on the criterion than those who were rejected.

Finally, in Figure 3, which shows the regression lines for the Number Facility test, the differentiation would be in the positive direction for all four groups, but less clear-cut for the two groups of white technicians than for the Negro technicians.

Two other aspects of these data should be examined in considering the questions of possible bias in selection and employment. One of these is the relative level of ability of each of the rating groups. For each rater-ratee ethnic combination, Table 2 show means and standard deviations of the aptitude tests, the Job Knowledge test, and Civil Service grade level. It can be seen from the table that on every test, mean scores for whites rated by Negroes were lower than those for whites rated by whites. For some tests these differences are not large, but they are all in the same direction. Thus it appears that the white medical technicians who had

Negro supervisors were less able than those who had white supervisors. For Negro technicians, on the other hand, mean scores did not vary systematically according to race of rater. (Neither the two white ratee groups nor the two Negro ratee groups were mutually exclusive, since some technicians were rated by both a Negro and a white supervisor. The elimination of these overlapping cases would, of course, increase the reported difference in the means.)

The second noteworthy aspect of these data has to do with the similarity of Civil Service salary grade level despite the differences in the test scores. For every test, the mean score for whites rated by whites ranked highest, that for whites rated by Negroes ranked second highest, and means for the two Negro ratee groups alternated (non-systematically) between the third and fourth ranks. Yet no such differences appeared in the means for Civil Service grade level. This discrepancy between what is indicated by the test scores and by the salary grade levels can be interpreted in a variety of ways, depending upon how one wishes to view the data. It might be interpreted as an example of criterion bias (with salary grade level serving as the criterion) working against the white group, in that superior job knowledge should rightfully be reflected in superior salary grade levels. On the other hand, it might be considered an illustration of test bias working against the Negro group, in that the grade level could be considered the realistic measure, with the discrepancy in test scores caused by bias in the content of the tests. These data alone do not provide us

with the means to choose between these two opposing interpretations. Rather, this example can be regarded as an illustration of the importance of the definitions of bias that are employed, as well as demonstrating the variety of interpretations that can be placed on a particular piece of objective data.

PART II - WEIGHTED TEST BATTERIES

The concern in this section was how a test battery which was selected on the basis of being most predictive for one ethnic group functioned when it was applied to another ethnic group.

There are two aspects to be considered in making such an evaluation. The first is the degree to which the predicted ratings parallel the actual ratings. This is measured by the multiple correlation coefficient or the cross-validated multiple correlation coefficient. The second is whether the predicted rating for an individual is higher or lower when the multiple regression equation for his own ethnic group is used rather than the multiple regression equation determined for the other ethnic group. In other words, is a Negro more or less likely to be employed if his test scores are weighted by a formula determined on an all-white validation group? Conversely, is a white more or less likely to be hired if his test scores are weighted by a formula determined on an all-Negro validation group?

Procedure

In these analyses, the criterion used was the average rating received from supervisors on a particular scale, regardless of whether the supervisor was Negro or white. It would have been desirable to have those analyses

by separate rater-ratee ethnic group combinations also, except that sizes of the two samples with Negro supervisors would have meant that too few degrees of freedom would have remained. Separate analyses, though, were done for Negro and white technicians. Stepwise multiple correlation coefficients were computed to give a best weighted prediction of each rating scale for the Negro and white samples separately. The weights determined on one sample were then applied to the other, to see what effect such "cross-ethnic cross-validation" would have.

Correlations

The correlations between predictor tests and supervisors' ratings of Negro and white subjects are given in Table 3. These are corrected for attenuation in the predictors as well as in the criterion scales. The present data are intended to show the validities potentially available in the predictors used, for predicting performance as medical technicians. The predictor tests were arbitrarily kept brief to allow for the collection of a variety of predictor, criterion, and background measures. (The correlations used subsequently in computing multiple correlations were not corrected for attenuation.)

Upon examining the pairs of validity coefficients in Table 3 column by column, it may be seen that validities for Negroes were higher than those for whites in all instances for the first, second, fourth, and sixth tests. The reverse was true on the fifth and eighth tests, where 17 of the 18 correlations were greater for whites. Thus, the general expectation that

pencil and paper tests are less valid for Negroes was certainly not borne out in the present instance.

The oft-voiced concern that school-oriented tests are less valid for Negroes than for whites also failed to hold for the population studied. Two of the four tests having consistently higher validities for Negroes than for whites are computational, the Subtraction-Multiplication and the Necessary Arithmetic tests. Another of the four is a test of vocabulary, and the last, Number Comparison, is a standard test of clerical ability. Tests that showed higher validities for whites, on the other hand, are the Fine Finger Dexterity test and the Picture-Number test. The latter is a test of short-term memory which would seem a likely candidate for a "culture-fair" test.

As has been indicated, the subjects of the study were incumbent medical technicians, rather than job applicants. On the other hand, there was not the usual problem of restriction of range due to testing, since the technicians studied had not been selected for their jobs on the basis of tests.

Multiple Correlation Coefficients

For each ethnic group, multiple correlations were computed for the best weighted combination of the nine experimental tests. These correlations are given in the first and third columns of Table 4, for whites and Negroes, respectively. In comparing the two sets of multiple correlations, note that for every rating scale, Negro weights applied to the Negro sample yielded a higher multiple correlation than did the white weights applied to whites.

Note further that the lowest multiple correlation for Negroes, .29 on the Overall rating, was exceeded by only two of the multiple R's for whites, .38 on Learning Ability and .36 on Flexibility. The conclusion is stregthened, then, that a battery of objective pencil and paper tests is indeed relevant for Negroes as well as whites in predicting rated job performance.

The comparatively high multiple correlations for Negroes could have come from the relatively culture-free tests, such as Picture-Number (testing short-term memory) or Finger Dexterity. Such was not the case, however. For nearly every scale, Subtraction-Multiplication and Necessary Arithmetic test scores were assigned the largest weights in the multiple correlations for Negroes. Picture-Number also appeared in several scales, but with a negative weight. For the white sample, Necessary Arithmetic again figured prominently, having the largest weight for five of the nine scales. Unlike the Negro multiple correlations, however, those for whites included sizable positive weightings on Finger Dexterity and Picture-Number scores.

Cross-Validation Coefficients

How well will a test battery selected for a white sample make generally valid predictions about Negroes? This question can be answered for the data just presented, by applying the weights determined on the white sample to obtain multiple correlations for the Negro sample. The cross-ethnic cross-validation coefficients resulting from doing this are given in the second column of Table 4. Similarly, the results of applying weights derived from Negro data to the white sample are given in the fourth column of the table.

When the weights determined on the white sample were applied to the Negro sample, five rating scales actually had higher multiples than they did for the white sample. This of course reflects the fact that the tests contributing to those multiples had higher validities for the Negroes than for the whites. Mulitples for three of the four remaining scales dropped only slightly. Thus, it appears that a battery selected for a white sample will make generally valid predictions among Negroes, as well. The converse was less true, as is apparent upon examination of the last two columns in Table 4. On most scales, there was considerable shrinkage in the multiple correlation when weights derived for the Negro sample were applied to the whites.

Predictions Resulting from Multiple Regression Equations

The multiple regression equations derived for each ethnic group were used to compute predicted criterion (rating) scores for three hypothetical individuals: (a) one whose test scores in the equations were precisely one standard deviation above the mean for his group, (b) one whose scores were at the mean for his group, and (c) one whose scores were one standard deviation below the mean for his group. The cross-ethnic cross-validation was achieved by also using the regression equations derived for the other ethnic group, with the same scores. These predicted criterion scores are shown in Table 5 for the Negro sample and Table 6 for the white sample.

In Table 5 it can be seen that on eight out of the nine regression equations, a Negro with high scores will fare better, that is, receive

higher predicted criterion scores, if the regression weights based on the Negro sample are used rather than weights based on the white sample. However, a Negro with low scores does better if the weights based on the white sample are used, in six out of the nine equations. Table 6 indicates that a white with high scores will do better for all nine equations if the weights based on the Negro sample are used. A white with low scores does better with the Negro weights in five out of the nine equations. These results of course reflect the earlier finding of higher validities (and hence, steeper regression slopes) for the regression equations based on the Negro sample, but higher mean scores (and thus a larger intercept constant) for the regression equations based on the white sample.

Summary and Conclusions

Several conclusions may be drawn from these analyses. One, the belief that pencil and paper tests are generally less valid for Negroes than for whites was not supported by the present study. Validity coefficients were generally somewhat higher for the Negro group than for the whites. In addition, there were consistently higher validities for Negroes than for whites on tests which might be considered "culture-bound", including Subtraction-Multiplication, Necessary Arithmetic, and Vocabulary; but there were higher validities for whites on tests one might assume to be "culture-free," including Finger Dexterity and Picture-Number.

Evidence that the pencil and paper tests were as valid for the Negro subjects as for the whites was even more pronounced when multiple correlations

were examined; and presumably "culture-bound" tests played as important a role compared to "culture-free" tests for Negroes as they did for whites. On all nine rating scales, multiple correlations computed for the Negro sample were greater than those computed for whites. Further, the more "culture-bound" tests such as Subtraction-Multiplication and Vocabulary were generally weighted more heavily for the Negro sample than for the white.

Cross-ethnic cross-validation of the weights derived from the white sample indicated that a test battery selected on this basis would be generally valid for Negroes, as well. The converse was less true. There was generally large attrition in multiple correlation when weights derived for the Negro sample were applied to whites.

However, the use of the multiple regression equations based on the Negro sample tended to favor whites. The use of the white regression equations would benefit Negroes with low test scores, but not those with high test scores.

The effect of using a single test for prediction depends on the particular rater-ratee ethnic group involved. Selecting the best predictor test for one rater-ratee ethnic combination may result in quite undesirable selection practices for the other rater-ratee ethnic combinations.

Table 1

Correlations of Aptitude Tests With Supervisors' Overall Ratings, for Ratings Grouped by Ethnic Combinations of Rater and Ratee

Tests	Negro Rated by Negro (N=approx. 50)	Negro Rated by White (N=approx. 148)	White Rated by Negro (N=approx. 41)	White Rated by White (N=approx. 276)
Subtraction-Multiplication	.41	.29	.14	.15
Vocabulary	.19	.08	-.16	.09
Hidden Figures	-.07	-.01	-.17	.07
Necessary Arithmetic	.17	.18	.19	.16
Finger Dexterity	-.02	.12	.20	.12
Number Comparison	.25	.19	.04	.05
Gestalt Completion	-.03	.05	.17	.11
Picture-Number	-.01	-.01	.03	.18
Paper Folding	-.04	.13	.27	.09

Table 2

Test Means and Standard Deviations by Rating Group

Measures	Negro Rated by Negro (N=approx. 50)		Negro Rated by White (N=approx. 148)		White Rated by Negro (N=approx. 41)		White Rated by White (N=approx. 276)	
	Mean	S. D.	Mean	S. D.	Mean	S. D.	Mean	S. D.
Predictor Test								
Subtraction-Multiplication	46.9	16.9	47.2	17.5	51.8	16.1	57.6	19.8
Vocabulary	19.1	7.1	19.1	7.6	24.9	9.7	25.7	9.0
Hidden Figures	6.0	3.4	5.8	4.1	7.5	4.4	7.8	4.9
Necessary Arithmetic	10.9	3.8	10.4	3.7	13.0	5.1	14.0	4.7
Finger Dexterity	22.2	15.4	22.7	15.0	24.5	14.3	29.2	15.7
Number Comparison	36.9	10.4	37.0	9.9	39.2	8.7	42.6	9.8
Gestalt Completion	10.3	4.9	10.3	5.3	11.8	6.0	12.0	5.5
Picture-Number	15.8	8.2	17.5	8.6	17.6	9.3	20.9	9.5
Paper Folding	6.4	3.1	6.9	3.3	8.9	3.8	9.1	3.7
Job Knowledge Test	32.6	10.8	31.0	10.6	34.2	11.5	36.0	9.0
Civil Service Grade Level	5.9	1.1	5.7	1.1	5.8	1.2	5.8	1.1

Table 3

Correlations Between Predictor Tests and Supervisors' Ratings on Selected Criterion Scales, Corrected for Attenuation in Criteria and Predictors

Rating Scale	Predictor Test 1. Subtr-Mult	2. Vocab	3. Hidden Figure	4. Nec Arith	5. Finger Dext	6. Number Compar	7. Gestalt Compl	8. Pict_ Number	9. Paper Folding
Flexibility	30	00	22	38	31	20	32	29	34
	48	16	06	46	19	22	20	-05	21
Planning	18	01	04	21	19	06	18	18	13
	51	16	05	34	14	24	10	-12	02
Interest	16	08	06	21	15	09	08	17	14
	40	14	05	27	05	10	04	-11	00
Learning Ability	30	09	17	40	32	21	25	27	38
	55	32	03	59	29	40	29	10	46
Job Knowledge	11	17	-01	16	12	-01	04	08	16
	41	27	10	49	11	24	19	-03	14
Technique	14	08	08	21	21	10	18	26	20
	37	21	06	35	10	23	11	-09	11
Low Need for Supervision	06	06	04	12	08	-01	04	12	08
	36	14	04	39	07	14	06	00	14
Communication	08	22	11	17	07	01	04	07	13
	32	31	-03	35	08	20	18	-08	18
Overall	15	07	06	20	14	05	13	19	14
	40	13	03	26	13	24	07	-03	13

Note -- In each pair of correlations, the upper and lower values are for the white and Negro samples, respectively.

Table 4

Multiple Correlation Coefficients and Cross-Ethnic Cross-Validation Coefficients for Predicting Supervisors' Ratings from Aptitude Test Scores

Rating Scale	White Weights		Negro Weights	
	White Sample (N = 297)	Negro Sample (N = 166)	Negro Sample (N = 166)	White Sample (N = 297)
1. Flexibility	36	34	41	24
2. Organization	19	18	36	11
3. Interest	15	17	32	07
4. Learning Ability	38	40	42	32
5. Job Knowledge	17	21	43	13
6. Technique	23	-01	35	07
7. Low Need for Supervision	11	04	33	05
8. Communication	17	21	34	15
9. Overall	16	17	29	13

Table 5

Ratings Predicted By Multiple Regression Equations

Negro Sample

Rating Scales	Test Scores One Standard Deviation Below the Mean		Test Scores At the Mean		Test Scores One Standard Deviation Above the Mean	
	Using Negro Weights	Using White Weights	Using Negro Weights	Using White Weights	Using Negro Weights	Using White Weights
Flexibility	4.83	4.17	5.46	4.92	6.08	5.69
Organization	5.29	5.40	5.75	5.75	6.21	6.09
Interest	5.25	5.49	5.57	5.70	5.88	5.91
Learning Ability	4.86	4.58	5.79	5.41	6.71	6.24
Job Knowledge	4.46	4.71	5.23	5.07	6.00	5.43
Technique	5.47	5.30	5.89	5.77	6.32	6.25
Need for Supervision	5.14	5.75	5.72	5.99	6.30	6.22
Communication	4.91	5.25	5.49	5.54	6.07	5.79
Overall	5.14	5.36	5.71	5.61	6.27	5.87

Table 6

Ratings Predicted By Multiple Regression Equations

White Sample

	Test Scores One Standard Deviation Below the Mean		Test Scores At the Mean		Test Scores One Standard Deviation Above the Mean	
Rating Scales	Using Negro Weights	Using White Weights	Using Negro Weights	Using White Weights	Using Negro Weights	Using White Weights
Flexibility	5.26	4.55	6.05	5.43	6.84	6.32
Organization	5.60	5.58	6.18	5.99	6.76	6.39
Interest	5.44	5.64	5.81	5.90	6.19	6.17
Learning Ability	5.40	5.00	6.51	5.96	7.63	6.92
Job Knowledge	5.03	4.90	6.01	5.30	6.99	5.71
Technique	5.79	5.49	6.34	6.01	6.90	6.54
Need for Supervision	5.54	5.85	6.27	6.10	7.00	6.36
Communication	5.35	5.45	6.09	5.75	6.83	6.06
Overall	5.40	5.54	6.05	5.87	6.69	6.19

FIGURE 1

Regression lines for predicting overall rating from short-term memory (Picture-Number) test score, for data grouped by race of rater and race of ratee.

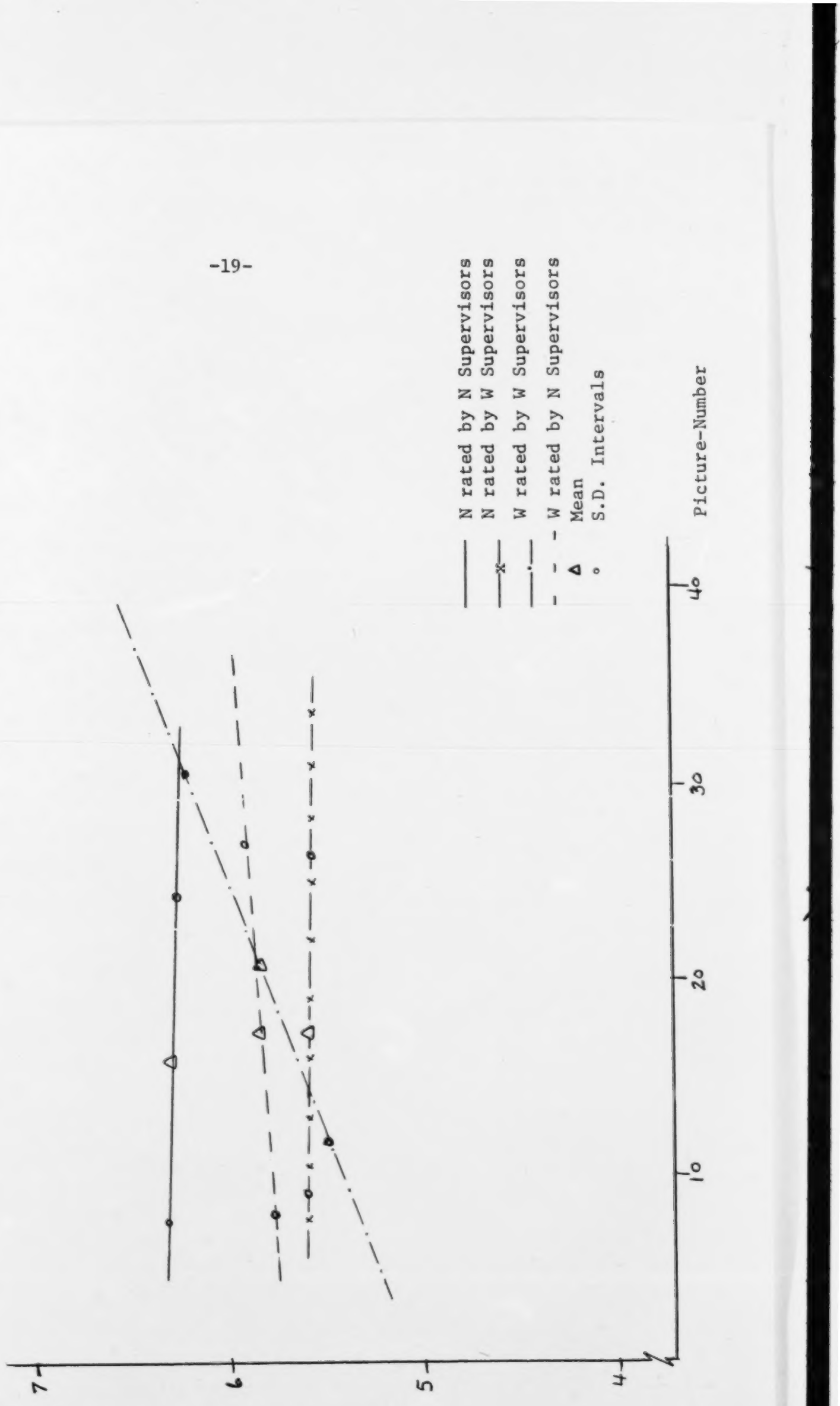


FIGURE 2

Regression lines for predicting overall rating from spatial visualization (Paper Folding) test score, for data grouped by race of rater and race of ratee.

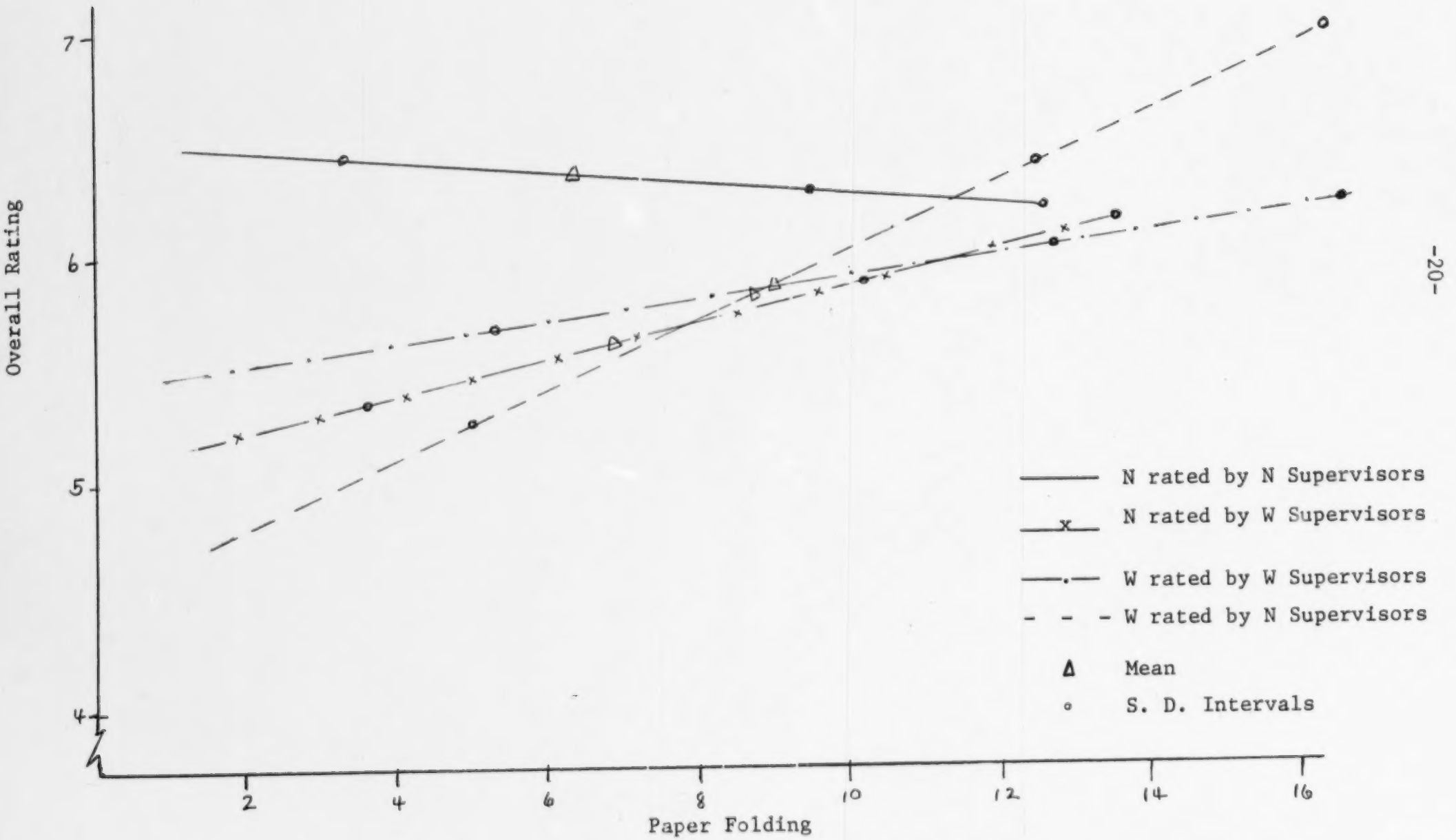
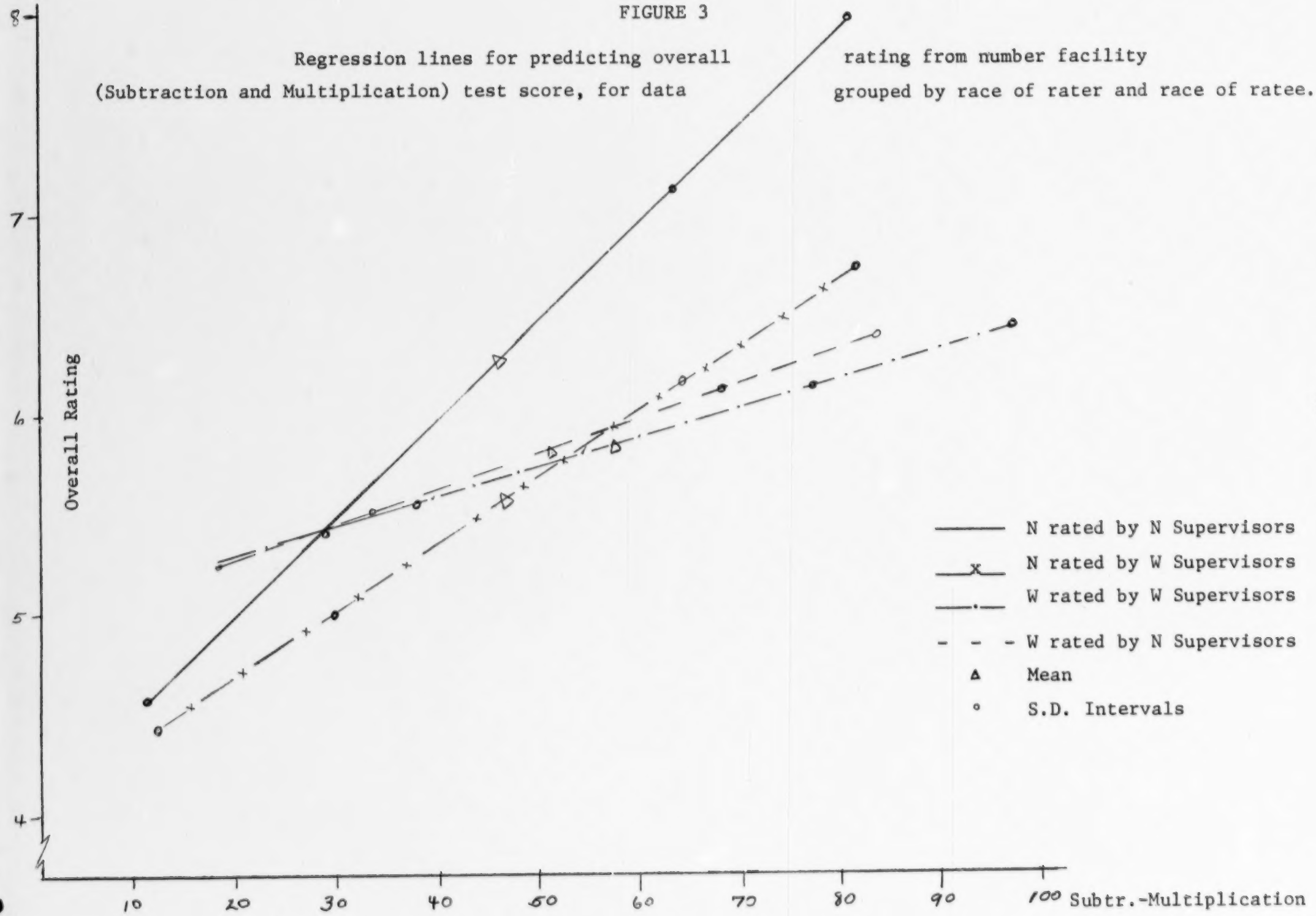


FIGURE 3

Regression lines for predicting overall rating from number facility (Subtraction and Multiplication) test score, for data grouped by race of rater and race of ratee.



Reference

Flaugher, R.L., Campbell, J.T., & Pike, L.W. Prediction of job performance for Negro and white Medical Technicians: Ethnic group membership as a moderator of supervisor's ratings. Project Report 69-5. Princeton, New Jersey: Educational Testing Service, 1969.